Outline

- The physics goal of this project is to classify Higgs to tau-tau decays (signal process) from background decays.
- The classification task aims to improve the statistical significance of the existence hypothesis \( H_0 \) and the signal process \( H \rightarrow \tau^+ \tau^- \) exists.
- The significance of the alternative hypothesis is quantified by the metric Approximate Median Significance \( (\sigma) \) which is another way of expressing a \( p \)-value.
- A significance of 5\( \sigma \) is required to claim a discovery of a new decay. The Higgs to tau-tau analysis has not yet achieved 5\( \sigma \) hence, the decay is unobserved in nature.
- Classifying signal from background is a challenging task since the classes completely overlap, the signal process is embedded within a dense background and is largely inseparable.

In this project we propose a meta algorithm for the binary classification task and measure its performance through the AMS \( (\sigma) \) metric.

Algorithm

1. Select \( K \) features \( x_1, \ldots, x_K \) at random from the training set \( D \).
2. Select \( K \) splits \( \{T_1, \ldots, T_K\} \), one per feature for the \( K \) features chosen in step 1; each \( T_i \) is selected at random from the range of the feature values \( \forall i = 1, \ldots, K \).
3. Rank the splits \( T_i \) by a criterion say \( Q \) which gives a score \( Q(D, T_i) \in \mathbb{R} \) for each split.
4. Return \( T^* = \max_{i=1,\ldots,K} Q(D, T_i) \).

Results

The tree (left) has a depth of 4 and the nodes \( [A, \ldots , E] \) represent leaves, these might not be pure (i.e. contain samples of one class). All samples end up in a leaf and are assigned a score equal to the fraction of events of the same class in that leaf.

For example, if leaf \( C \) has 30 background events and 2 signal events, the signal events in leaf \( C \) will have a score of 1/16 and background events will have a score of 15/16.

References